This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

THIS PAGE BLANK (USPTO)

09/868554
PCT/JP99/07050

#4

REC'D 10 MAR 2000
WIPO PCT

EJN

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

24.01.00

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

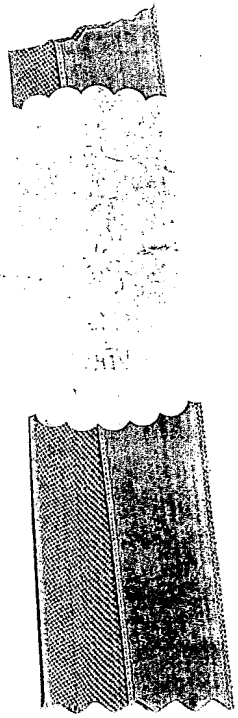
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年　８月２５日

出願番号
Application Number: 平成１１年特許願第２３８０５３号

出願人
Applicant (s): 松下電器産業株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)



２０００年　２月２５日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号　出証特２０００－３００９６４４

【書類名】　特許願

【整理番号】　2033811021

【提出日】　平成11年 8月25日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 15/403

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　近藤　堅司

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　今川　太郎

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　松川　善彦

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　目片　強司

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100097445

【弁理士】

【氏名又は名称】　岩橋　文雄

【選任した代理人】

【識別番号】　100103355

【弁理士】

【氏名又は名称】　坂口　智康

【選任した代理人】

【識別番号】　100109667

【弁理士】

【氏名又は名称】　内藤　浩樹

【先の出願に基づく優先権主張】

【出願番号】　平成10年特許願第355657号

【出願日】　平成10年12月15日

【手数料の表示】

【予納台帳番号】　011305

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9809938

【プルーフの要否】　不要

【書類名】　明細書

【発明の名称】　検索処理方法

【特許請求の範囲】

【請求項１】　１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列データを検索することを特徴とする検索処理方法。

【請求項２】　文書データ群に文字要素同士の接続関係を複数通り保持しておき、前記文書データ群の中から指定した文字列とあらかじめ定めた関係を満たす文字要素列データを検索することを特徴とする検索処理方法。

【請求項３】　接続する文字要素の位置を複数保持しておくことを特徴とする請求項２記載の検索処理方法。

【請求項４】　接続する文字要素を複数保持しておくことを特徴とする請求項２記載の検索処理方法。

【請求項５】　指定した文字列を構成する文字要素数およびまたは文字種に基づいて検索手段を切り替えることを特徴とする１～４いずれかに記載の検索処理方法。

【請求項６】　指定した文字列を構成する文字要素のうち文書データ群に含まれる文字要素と予め定めた関係を満たす文字要素の状態と、検索結果が検索対象の文字列である可能性を表す値を対応づけたテーブルを用いて、検索した文字要素列データの正当性を判定することを特徴とする請求項５に記載の検索処理方法。

【請求項７】　前記状態は、前記指定した文字列を構成する文字要素のうち文書データ群に含まれる文字要素と前記関係を満たした文字要素数であることを特徴とする請求項６に記載の検索処理方法。

【請求項８】　前記状態は、前記指定した文字列を構成する文字要素数、前記指定した文字列に対する前記関係を満たした要素の位置、前記要素の文字種、のうちの少なくとも１つ以上と、前記指定した文字列を構成する文字要素のうち文書データ群に含まれる文字要素と前記関係を満たした文字要素数との組み合わせ

であることを特徴とする請求項６に記載の検索処理方法。

【請求項９】　判定した前記正当性を用いて、検索手段を切り替えることを特徴とする請求項６～８いずれかに記載の検索処理方法。

【請求項１０】　指定した文字列を構成する文字要素群を複数の文字要素群に分割し、それぞれの文字要素群に対し、文書データ群の中から予め定めた関係を満たす文字要素列データを検索することを特徴とする請求項６～９のいずれかに記載の検索処理方法。

【請求項１１】　指定した文字列を構成する文字要素群を、複数の単語に分割することを特徴とする請求項１０に記載の検索処理方法。

【請求項１２】　単語中の部分文字列も単語とみなすことを特徴とする請求項１１に記載の検索処理方法。

【請求項１３】　文書データ群に検索パラメータを付加しておき、前記パラメータにを用いて検索手段を切り替えることを特徴とする１～１２いずれかに記載の検索処理方法。

【請求項１４】　文書データ群中の文字要素毎に検索パラメータを付加しておくことを特徴とする請求項１３記載の検索処理方法。

【請求項１５】　指定した文字列を構成する文字要素と文書データ群を構成する文字要素との距離があらかじめ定めた基準以下のデータを検索することを特徴とする請求項１～１４いずれかに記載の検索処理方法。

【請求項１６】　検索パラメータを用いて、距離の基準を決定することを特徴とする請求項１５記載の検索処理方法。

【請求項１７】　指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列データを文字要素間の距離を表現したテーブルを用いて検索することを特徴とする請求項１～１６いずれかに記載の検索処理方法。

【請求項１８】　複数のテーブルを有し、テーブルを切り替えて検索を行うことを特徴とする請求項１７記載の検索処理方法。

【請求項１９】　指定した文字列を構成する文字要素数およびまたは文字種およびまたは検索パラメータに基づいてテーブルを切り替えることを特徴とする請求項１８記載の検索処理方法。

【請求項２０】　発生頻度パラメータをテーブルに付加しておき、前記パラメータを用いて指定した文字要素に対して指定した距離に位置する文字要素の発生頻度を算出し、前記発生頻度に基づいて検索を行うことを特徴とする請求項１７～１９いずれかに記載の検索処理方法。

【請求項２１】　発生頻度があらかじめ定めた基準以上のデータを検索することを特徴とする請求項２０記載の検索処理方法。

【請求項２２】　発生頻度に基づいて他の文字要素に対する検索手段を切り替えることを特徴とする請求項２０記載の検索処理方法。

【請求項２３】　指定した文字列を構成する文字要素のうち、あらかじめ定めた基準の距離又は発生頻度を満たす文字要素の割合に基づいて検索手段を切りかえることを特徴とする請求項２０～２２いずれかに記載の検索処理方法。

【請求項２４】　指定した文字列を構成する文字要素をテーブル中の他の文字要素に置き換えて文書データ群から検索することを特徴とする請求項１７～１９いずれかに記載の検索処理方法。

【請求項２５】　文書データ群中の文字要素をテーブル中の他の文字要素にあらかじめ置き換えたデータを文書データ群に付加しておくことを特徴とする請求項１７～１９または２４いずれかに記載の検索処理方法。

【請求項２６】　検索結果に基づいて動作を決定することを特徴とする請求項１～２５いずれかに記載の検索処理方法。

【請求項２７】　指定した文字列を構成する文字要素を検出した位置または位置関係に基づいて動作を決定することを特徴とする請求項２６記載の検索処理方法。

【請求項２８】　文書データ群中における検索すべき文字要素列の有無または位置に基づいたコマンドを出力することを特徴とする実施例２６記載の検索処理方法。

【請求項２９】　指定した文字列と検索すべき文字要素列との関係を異なる手段で再度判断することを特徴とする請求項１～２８いずれかに記載の検索処理方法。

【請求項３０】　請求項１～２９のいづれか一つの請求項に記載の各手続きの

全部または一部の手続きをコンピュータに実行実現するためのプログラムを格納したことを特徴とする情報記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は文字認識した文書を含むデータから指定した文字列に基づき、前記文字列に関連したデータを検索し、活用する技術に関するものである。

【０００２】

【従来の技術】

従来の文字認識した文書を含むデータから指定した文字列に基づいて関連したデータを検索する技術としては特開平７-152774「文書検索方法および装置」がある。

【０００３】

従来の方法の例を図２３を用いて説明する。図２３は紙に記された文書と前記文書を文字認識した場合の結果を示している。通常文字認識においては、紙面に印字された文字のかすれや傾き、字体、文字サイズなどの影響で認識誤りを生じ得る。図２３においてはオリジナルの文書における「本」という文字が「木」という字に誤って認識されている。

【０００４】

また、オリジナルの文書における「口」という文字が「区」という文字に誤認識されている。ここで、「日本」という文字列を検索する場合を考える。このとき、表１に示すような誤認識文字の表を用いる。誤認識文字の表はあらかじめ、文字認識によって間違われやすい文字を並べた表である。表１においては、「本」という文字は「木、大、太、才」に誤って認識されやすく、「口」という文字は「口(記号の四角形)、回、円、々」に間違われやすいということを示している。

【０００５】

【表１】

対象文字	誤認識文字
本	木、大、太、才
口	口、回、円、々

【０００６】

「日本」を検索する場合、文字認識された文書より「日本」という文字列を検索すると同時に誤認識文字の表を用いて「日木」、「日大」、「日太」、「日才」という文字列を生成し、これらの文字列も「日本」同様に検索することで「日本」が誤認識された「日木」の部分を望ましく検索できるようになる。

【０００７】

【発明が解決しようとする課題】

しかしながら上記従来の手法ではあらかじめ、誤認識しやすい文字を用意しておくために、誤りの少ない文書データを検索する時には余分な文字候補を用いた余分な検索処理が行われ、また逆に誤りの多い文書データではあらかじめ用意した誤認識文字表に含まれる文字以外の誤認識には対応できない場合が発生するという課題を有していた。

【０００８】

例えば、図２３において「人口」という文字列を検索したい場合、誤認識文字の表を用いて「人口(記号の四角形)」、「人回」、「人円」、「人々」を同時に検索するが、誤認識文字の表に存在しない誤り（「口」を「区」に間違う）が発生した場合である「人区」（本来は「人口」）は検索が不可能であった。

【０００９】

また、一般の文書を文字認識したデータを検索する場合、文字認識時の文書レイアウトの判断誤り（縦書きと横書きの誤判断、改行後の次行への接続の判断誤り、段落から段落への接続の判断誤りなど）が起こり得るが、上記手法ではレイアウトの誤りに対しては対応できないという課題を有していた。

【００１０】

例えば、図２４のようなレイアウトの文書を文字認識する場合を考える。図２４において段落の正しい順番は、右上の段落、左上の段落、右下の段落、左下の段落である。しかしながら、文字認識の過程において、段落の順番を誤って判断し、右上の段落の次に右下の段落が接続すると判断する場合が起こり得る。ここで「日本の人口」という文字列を検索したい場合、誤認識表などを用いて個々の文字について望ましい検索が可能であっても、段落の接続が誤っている場合、図２５のように「・・・日本のする傾向・・・」という文書として扱われるため、「日本の人口」という文字列は検索できない。

【００１１】

そこで、本発明は上記従来の課題を鑑み、検索したい文字列を複数の要素（文字または文字片または文字列または文字片と文字との組合わせ）に分け、それぞれを独立に検索することで、複数行にわたる文字列の検索を行う際に、レイアウトの誤認識が含まれていても検索を可能とする。

【００１２】

また、文字要素同士の関係をテーブルとしてを保持しておくことで、許容できる誤りの度合いが可変かつ高速な検索を可能とする。

【００１３】

【課題を解決するための手段】

本発明は１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、

文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索することを特徴とする検索処理方法である。

【００１４】

また、文字要素間の距離を表現したテーブルを用いて、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索すること検索処理方法である。

【００１５】

【発明の実施の形態】

（第１の実施の形態）

以下、本発明の実施の形態について図面を参照して説明する。図１は本発明の第１の実施の形態で用いる文字要素同士の距離を示す距離テーブルの一例を示すものである。ここで、文字要素とは「亜」のような文字そのものや、「00」のような複数の文字の集まったもの、図２のような文字を構成する文字片や、図３のような文字片と文字とが集まったものなどを示す。また、文字には「)」や「◎」のような記号も含める。

【００１６】

距離テーブルは文字要素間の近い・遠いの関係を数値で表現したものである。図１では文字要素「亜」と文字要素「唖」との距離が１０、文字要素「亜」と文字要素「00」との距離が１７２であることを示しており、文字要素「亜」は文字要素「00」よりも文字要素「唖」に距離が近いことを示している。他の文字要素についても同様に文字要素間の距離を定義しておく。

【００１７】

距離の定義に関しては、特定の文字認識システムの入出力関係、各文字要素の形状を特徴量数値で表現した場合の特徴量空間内でのユークリッド距離などを用いることができる。

【００１８】

また、距離テーブルは文字要素間の距離を表現していれば、必ずしも図１のような格子状の表の形態をとる必要はなく、文字要素ごとに他の文字要素との距離を距離の近い順番に保持していても良く、また順番そのものを距離と扱うことも可能である。

【００１９】

第１の実施の形態について、文字要素間距離テーブルを用いた文字列の検索について、その手続きを説明する。ここで、図７のように本来「・・・日本の人口構成は・・・」である文書を文字認識して得られた「・・・日木の人区構成は・・・」という認識結果（文書データ、あらかじめ記憶媒体などに保持しておく）から文字列を検索することを考える。

【００２０】

通常、文字認識技術を用いた場合様々な要因で誤りが生じる。この場合、「本

」という字が誤って「木」に認識され、「口」という文字が誤って「区」という文字に誤って認識されている。ここで、「日本」という文字列を指定して、本来「日本」が存在した場所を図７の認識結果の中から検索することを考える。まず指定した文字列「日本」を構成する文字要素として「日」について文字要素間距離テーブルを参照し、距離があらかじめ定めた値（例えば１５０など）よりも小さい文字（例えば「日」については「日」と「目」）を図７の認識結果から検索する。

【００２１】

この場合、認識結果の中の「日」という文字要素を結果として検出する。次に指定した文字列「日本」を構成する次の文字要素として「本」について文字要素間距離テーブルを参照し、距離があらかじめ定めた値（同上の１５０）よりも小さい文字（「本」については例えば「本」と「木」と「大」）が「日」を検出した位置の次の位置の文字要素「木」と一致するかを図７の認識結果から判断する。この場合、文字要素「木」が一致することから、指定した文字列「日本」に対して図７の認識結果中の「日木」が検出できる。これによって本来「日本」という文字列が誤って「日木」と認識された場合にでも、本来の文字列の位置を検索することが可能となる。実際には指定した文字要素列を検出した場合、検出した文字要素列のみならず、検出した文字要素列を含む前後の文書の認識結果も合せて検索者に提示したり、文字認識を行ったオリジナルの文書を画像イメージとして別途文書データに保持しておき、対応する画像イメージを検索者に提示することで、例え文字認識の結果が一部誤っていた場合にでも、人間が再度判断することで、検索者が必要とする情報を得ることが可能となる。

【００２２】

また、指定した文字要素列を検出した際には、前後の文書の提示以外に文書のタイトルや要約を表示しても良い。この場合、少ない表示スペースで検索結果を把握することが可能となる。また、表示以外に音声を用いて前後の文章やタイトル、要約を出力することで、表示領域の少ない端末にも対応することができる。また、出力は通信路（ネットワーク）を経由して出力しても良い。また、帯域が狭い通信路を経由する場合には、検索結果の画像イメージを最初から表示するの

ではなく、前後の文書の認識結果やタイトル・要約のみを最初に表示し、検索者が別途指示することで情報量の多い画像イメージの表示を行うことで、検索時間や閲覧時間を節約することが可能となる。

【００２３】

更に、指定した文字要素列が検出できた際に、検出した情報を提示するのではなく、機器への新たな命令（コマンド）を発行しても良い。例えば、リアルタイムにカメラなどから得られる画像に対して検索を行い、特定の文字要素列を検出した場合（例えば「レストラン」）に撮像を行う機器に対して映像をメモリに記録するコマンドを発行したり（レストランの映像を集めることが可能となる）、特定の文字要素列を検索した場合には、文字要素列を含む画像をプリンターへ印刷するコマンドをプリンタへ発行することや、文字要素列を含む画像の情報を通信網（ネットワーク）を通して複数の宛名に配信することなどが考えられる。

【００２４】

なお上記の場合、文字要素間距離について１５０という値をあらかじめ定めたが、この値は可変であり、最初に小さい値を設定して検索を行い、検索できない場合に順次大きな値に再設定して検索しても良い。これは最初に小さい距離に相当する値を設定することで、文字認識について誤りをあまり許容しない状態で検索を行い、順次文字認識の誤りを許容して検索することに相当する。よって、最初から文字認識の誤りを大きく許容することで、関係の無い余分な文字要素列の検出が発生することを未然に防ぐことが可能となる。

【００２５】

また、文字認識結果を保持するデータ（文書データ）に文字認識時の信頼度(又は尤度、確度、確からしさなど)を検索パラメータとして保持しておくことで、前記検索パラメータに応じて検索に用いる距離の基準値を適切な値に設定することが可能となる（図９）。図９では文書データ中の各文字の認識結果に対して認識時の信頼度を検索パラメータとして付与している。ここで信頼度の値は０から１までの値をとり、値が大きいほど認識結果が確からしいと判断する。ここで「人口構成」という文字要素列を検索する場合を考える。「人」・「口」・「構」・「成」と他の文字との距離関係の一部を図１０の通りとする。図９では認識

結果の「人」・「成」の信頼度は0.9と高いので文字要素間距離テーブルで許容する距離を比較的小さな値１０と設定し、逆に認識結果の「区」のように信頼度の低い（0.4）場合には許容する距離を６０と大きい値にする。誤認識している「区」については信頼度が低いために距離の基準を低くすることで図１０のテーブルから「区」との距離が５０である「口」の検索対象となる。その結果、「人口構成」という文字要素列を指定して、誤認識文字を含む「人区構成」という文字要素列を検索することが可能となる。

【００２６】

このように認識の信頼度の低い文字要素または文書については、検索時の文字要素間距離テーブルで許容する距離を大き目に設定し、逆に認識結果の信頼度が高い場合には、許容する距離を小さ目に設定することで、関係の無い余分な文字列要素の検出を抑えることが可能となる。なお、信頼度の値と文字要素間距離テーブルで許容する距離との対応関係はあらかじめ定めておく。また、更に認識結果の信頼度が特に低い場合には、全ての文字の可能性を考慮したりするなど検索手段を切り替えてもよい。検索パラメータ（信頼度）の付与については文書毎や文字要素毎に付与してもよい。また、文字認識の信頼度としては、文字認識を行う際の認識システム（例えばニューラルネットワークなど）の出力や認識候補の数などを用いることが可能である。

【００２７】

ここでは、指定した文字列を構成する文字要素の先頭「日」から順番に一文字ずつ検索したが、異なる順番でも良い。特に一般的な文書中に出現する頻度を考慮し、指定した文字列を構成する文字要素のうち、一般的に文書中に出現する頻度の低い文字要素から検索することで、余分な検索手続きを減らすことが可能となり、検索速度を速めることが可能となる。

【００２８】

なお、上記の例では文書データを記憶媒体（メモリや磁気ディスク、光ディスクなど）にあらかじめ保持してくことを想定したが、画像入力機器（スキャナ、デジタルカメラ、ビデオカメラなど）から入力した画像情報を逐次文字認識して得られるリアルタイムの情報について同様の検索を行っても良い。

【００２９】

このように、文字要素間の距離テーブルを用いて文字要素列の検索を行うことで、指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要素列を文書データから検索することが可能となる。距離テーブルを用いることで、複雑な距離計算などを行う必要もなく高速な検索が可能となる。また、距離テーブルを用いることで、誤認識の許容度合いを都度適切な値に設定することが出来、効率の良い検索が可能となる。更に、検索パラメータを文書データに付与しておくことで、文書データや文字列要素に合わせた検索のための基準の選択や検索手法の切り替えが可能となり、検索の精度を向上させることが可能となる。

【００３０】

（第２の実施の形態）

次に、本発明の第２の実施の形態について説明する。第２の実施の形態は複数の文字要素間距離テーブルを用いた例である。検索の基本的手続きは第１の実施の形態と同様である。第２の実施の形態においては、複数の異なる文字要素間距離テーブルを用いて検索を行う。複数の文字要素間距離テーブルとしては、複数種類の文字認識システムそれぞれに対応したテーブルや、文字種（漢字、アルファベット、ギリシャ文字、カタカナなど）ごとに対応したテーブル、フォント種（ゴシック体のテーブル[図１]、明朝体のテーブル[図８]など）ごとに対応したテーブルなどをあらかじめ用意し、検索する文書データに応じて用いるテーブルを切り替える。例えば、文字認識して得た文書データについてあらかじめ、そこに含まれる文字の字種やフォントの種類、用いた認識システムの種類などを文書データに検索パラメータとして別途保持しておくことで、検索の際に適切なテーブルを選択して用いることが可能となり、検索の精度と速度を向上させることが可能となる。フォントごとにテーブルを切り替える場合、文字認識して得た文書データにあらかじめ文字ごとに明朝体に近いかゴシック体に近いかという情報を検索パラメータとして付与しておき、ゴシック体の文字から成る文書データより文字要素列を検索する場合には図１のようなテーブルを用い、明朝体の文字から成る文書データより文字要素列を検索する場合には図８のようなテーブルを用いる。フォントの種類の情報については文字認識を行う際に同時にフォントの種類

を認識することなどにより得ることができる。

【００３１】

また、同一の文書データに対しても複数のテーブルを切り替えても良い。この場合、特定のテーブルを用いて検索できなかった文書データについて再度、検索したい文字要素列の有無を異なる尺度で検証し、検索の精度を向上させることが可能となる。更に、テーブルを用いた検索の結果に対して、前記検索で得られた文字要素列の位置に対応する文書の画像イメージを用いて再度高精度な文字認識を行っても良い。これにより、テーブルを用いて高速な粗検索を行い候補を絞った後に、高精度な文字認識（一般的に処理時間がかかる）を用いて検索対象を確定することが可能となり、検索精度と検索速度の両立が可能となる。特に文字数の少ない検索文字列（２文字単語など）を検索する場合、類似した文字列が偶然検索される可能性が高い。よってこのような場合にも指定した検索文字列の文字用素数に基づいて異なるテーブルを用いて再検証したり、高精度な文字認識を併用したりすることで、処理時間を必要以上に増さずに、精度の高い検索が可能となる。

【００３２】

また、指定した検索文字列の字種に応じてテーブルを切り替えてもよい。例えば、検索文字列および文書データがアルファベットのみを含む場合には、アルファベット用のテーブルを用いることで、余分な検索処理を省くことが可能となる。

【００３３】

なお、上記例では１文字単位の文字認識の誤りを補う為に文字要素間距離テーブルを用いていたが、文字認識を行う際には複数の文字要素を単一の文字として扱った結果生じる誤認識（図４：２つの「木（き）」を「林（はやし）」と認識、２つの「０（ゼロ）」を「∞（無限大）」と認識）や、逆に単一の文字を複数の文字として扱った結果生じる誤認識（図５：「川」を３つの「１（いち）」と認識、「い」を「し」と「１（いち）」と認識）が存在する。このような場合にも、文字要素距離テーブルに「木（き）」２文字と「林」との距離、「０（ゼロ）」２文字と「∞（無限大）」との距離、「川」と３つの「１（いち）」、「い

」と「し」、「１（いち）」との距離がそれぞれ小さいことを保持しておく。図６に図４や図５の例を含む文字要素間距離テーブルの一例を示す。図６では「木（き）」２文字と「林」との距離、「０（ゼロ）」２文字と「∞（無限大）」との距離、「川」と３つの「１（いち）」、「い」と「し」、「１（いち）」との距離はそれぞれ１３以下で、他の組み合わせの場合（距離９８以上）よりも小さい値になっている。

【００３４】

ここで、検索したい文字要素列として「１００」を指定した時に、誤りの度合いの基準となる距離を５０とすると、「１」を検索した後、「０」を２つ検索すると同時に「∞」も検索することで、「１００」が「１∞」と誤認識された文字要素列を検索することが可能となる。同様に、検索したい文字要素列として「いろり」を指定した時に、誤認識された「し１ろり」という文字要素列を文書データから検出することなども可能となる。

【００３５】

更に、かな漢字変換等の誤りによってオリジナルの文書自体に文章として誤った表現が含まれる場合（「納める」を「収める」と表現）や、複数の送り仮名付け方が存在する場合（「変る」を「変わる」）や、漢字表記した言葉をひらがなで検索しようとする場合（「切磋」を「せっさ」で検索）や、類義語で検索しようとした場合（「価格」を「定価」で検索）や、異なる言語に対して検索しようとした場合（「history」を「歴史」で検索する場合）についても、文字要素間距離テーブルにおいてそれぞれ「収」と「納」との距離、「変わ」と「変」との距離、「切磋」と「せっさ」との距離、「価格」と「定価」のと距離、「history」と「歴史」との距離を小さな値として定義しておくことで検索することが可能となる。

【００３６】

（第３の実施の形態）

次に、本発明の第３の実施の形態について説明する。第３の実施の形態は文字要素間距離テーブルに文字要素間距離に加えて文字要素の発生頻度を付与しておき検索に用いる例である。図１３は発生頻度を付与したテーブルの例である。図

１３では「下」に対して「T」が距離１０では発生頻度0.2、距離２０では発生頻度0.6、距離３０では発生頻度0.2であることを示す。他にも図１４や図１５のように誤認識の発生頻度の分布を表現してもよい。図１４では正規分布を仮定し、平均の距離と分散をテーブルに保持している。例えば図１４においては、「下」に対してに「T」は距離２０を中心とし分散１０の正規分布に位置していることを表している。また、図１５では一様分布を仮定し、分布の最短距離と最大距離をテーブルに保持している。例えば、「下」に対して「F」は距離５０から７０までに一様分布し、「ト」は距離６３～１２２まで一様分布している。従って、距離６３から７０までの間には「F」と「ト」が重複して分布するのでそれぞれの発生頻度は0.5であると判断する。このように、図１３、１４、１５のようにテーブルに発生頻度を付与することで、文字要素と文字要素の距離を固定した一つの値でなく、幅を持たせると同時に、発生頻度に応じた検索が可能となる。

【００３７】

ここで、発生頻度を付与したテーブルを用いた検索の例を図１１を用いて説明する。基本的な検索の手続きは第１、二の実施の形態と同様であるが、信頼度に応じて文字間距離の許容する距離を決定した後、対応する距離での文字の発生頻度をテーブルから算出する。図Ｂではテーブルを用いて各文字要素について発生頻度を算出した結果を示している。図１１の文書データ中の「人」(信頼度0.9)に対して許容する距離を10(信頼度と距離および発生頻度との関係はあらかじめ定めておく)とすると、検索語の「人」に対して文書データ中の「人」が対応する割合(発生頻度)は0.9となっている。同様に、文書データ中の「区」(信頼度0.4)に対して許容する距離を60とすると、検索語の「口」に対して文書データ中の「区」が対応する割合(発生頻度)は0.1となる。「構」・「成」についても同様に発生頻度を0.9となる。ここで、指定した検索文字列「人口構成」を構成する各文字要素について文書データ中の文字要素列と対応する割合が平均0.7（=(0.9+0.1+0.9+0.9)／4）となる。あらかじめ検索の基準として一致する割合を例えば平均値で0.5以上としておくことで、上記の誤認識文字列「人区構成」を検索語「人口構成」から検索することが可能となる。更に、検索の結果を表示する際に、一致する割合に応じて表示を切り替えることも可能である。例えば、一致する

割合の高さに応じて文書画像中の対応する位置に強調表示（輝度や色、点滅などによる強調）を行うことで、検索者が一致割合を視覚的に確認することが容易となる。

【００３８】

また、上記の例では検索語を構成する文字要素の一致する割合の平均値を検索の基準としたが、最低値を基準としても良いし、、高い一致割合(例えば0.8以上)の文字が文字要素列全体のある割合以上(例えば半分以上)を占めている場合を基準としてもよい。更に、高い一致割合(例えば0.8以上)の文字要素が文字要素列全体のある割合以上(例えば２／３以上)あれば、残りの一致割合の低い文字要素に対しては検索の基準をゆるくして検出しやすいように変更してもよい。例えば、図１２のように「人口構成」が誤認識された「人同構成」という文字列（文書データ）を検索語「人口構成」で検索することを考える。ここで、「口」が誤認識された「同」は認識の信頼度が0.3で、許容する距離が80となり、「口」に一致する割合は0.0となっている。このままでは、「同」と「口」は全く一致しない扱いとなるが、検索語を構成する他の文字要素「人」・「構」・「成」については一致割合が高い（0.9）ので「同」に対しては許容する距離を80よりも大きく(例えば120)することで検出できるようになる（図１０の距離テーブルの場合）。

【００３９】

このように、テーブルに文字要素間距離に加えて、発生頻度の情報を付与しておくことで、同じ距離でも発生頻度に応じて検索の基準や手続きを切り替えることが可能となり、より精度の高い検索が可能となる。

【００４０】

（第４の実施の形態）

次に、本発明の第４の実施の形態について説明する。第４の実施の形態は文字要素間距離テーブルを用いて指定した文字要素列をあらかじめ複数の文字要素列に置き換えて検索を行う例である。図７の認識結果（文書データ）から指定した文字要素列として「日本」を検索する場合を考える。最初に文字要素列「日本」を構成する文字要素「日」と「本」に分け、それぞれの文字要素について文字要

素間距離テーブルを参照し、距離があらかじめ定めた値（例えば１５０など）よりも小さい文字（例えば「日」については「日」と「目」、「本」については例えば「本」と「木」と「大」）同士を組み合せて新たな文字要素列「日本」、「目本」、「日木」、「目木」、「日大」、「目大」を生成する。次に前記生成した文字要素列それぞれについて図７の認識結果（文書データ）より検索を行う。この場合、生成した「日木」がオリジナルの文章で「日本」が存在する位置において検出でき、望ましい検索結果を得ることが出来る。

【００４１】

ここで、検索した結果何も検出されない場合には、再度距離の基準値を大きくし（例えば２００に設定する）、新たにより多くの文字要素列を生成して同様な検索することで、距離テーブルで許容する距離が１５０の時には検出できなかった認識誤りを検出することも可能である。

【００４２】

このように、文字要素間距離テーブルを用いて指定した文字要素列を誤りの可能性のある複数の文字要素列に置き換えて検索することでも第１の実施の形態と同様に指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要素列を文書データから検索することが可能となる。距離テーブルを用いることで、複雑な距離計算などを逐次行う必要がなく高速な検索が可能となる。また、距離テーブルを用いることで、誤認識の許容度合いを都度適切な値に設定することが出来、効率の良い検索が可能となる。

【００４３】

（第５の実施の形態）

次に、本発明の第５の実施の形態について説明する。第５の実施の形態は文字要素間距離テーブルを用いて認識結果の文書データ中の文字要素に他の複数の文字要素を付加した後検索を行う例である。図７の認識結果（文書データ）から指定した文字要素列を検索する場合を考える。ここで、あらかじめ文字要素間距離テーブルを参照し、認識結果の文書データの各文字要素について距離があらかじめ定めた値（例えば１５０など）よりも小さい文字（例えば「日」については「日」と「目」、「木」については「本」と「大」など）を付加しておく（図１６

）。次に指定した文字要素列として「日本」を検索する場合には、「日本」を「日」と「本」に分け、最初に「日」を検索する。この場合図１６の文書データの１列目で「日」が検出できる。次に「日」を検出した位置の次の文字要素（「木」、「本」、「大」）の中に「本」が存在するかを判断し、「本」が含まれているので「日本」という文字要素列が検出できたとする。文字要素列の要素数が３つ以上の時にも同様な手続きで検出を行い、全ての文字要素が連続して検出できた場合に指定した文字要素列を検出できたと判断する。

【００４４】

このように、本実施の形態においては文字要素間距離テーブルを用いて認識結果の文書データ中の文字要素にあらかじめ複数の文字要素を付加しておくことで第１の実施の形態と同様に指定した文字要素列が誤認識で他の文字要素に置き換わったような文字要素列を文書データから検索することが可能となる。また、あらかじめ文書データ中の各文字要素に複数の異なる文字要素を付加しておくことで、検索時に距離テーブルを参照する手続きを省くことが可能となる。

【００４５】

なお、上記の形態に加えて、第１～第３の実施の形態のように検索の過程で距離テーブルを用いる形態や、第４の実施の形態のように指定した文字要素列を距離テーブルを用いてあらかじめ他の文字要素列に置き換える形態を併用して実施することも可能である。

【００４６】

（第６の実施の形態）

次に、本発明の第６の実施の形態について説明する。図１７のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。図１７において文章の正しい順序は段落Ａ、段落Ｂ、段落Ｃ、段落Ｄの順番である（一般的に様々なレイアウトを想定した場合、自動的に段落同士の接続の正しい順番を決めることは難しく、接続の誤りが発生し得る）。ここで、文字認識した結果を各段落ごとに付与した固有の番号（図１７ではＡ，Ｂ，Ｃ，Ｄ）と共に格納しておく（図１８）。このとき各段落について次に接続する可能性のある段落の番号も合せて記録しておく。図１８の場合、段落Ａの

後に接続する可能性のある段落はＢとＣであることを意味している。

【００４７】

ここで、接続の可能性のある段落の決め方としては、オリジナルの画像を文字認識する際に各段落同士のそれぞれの位置関係を参照することで決める。例えば、縦書きの場合、ある段落Ｘの次に続く段落としては段落Ｘよりも下に位置するか、左に位置する段落を接続の可能性のある段落とする。更に、文字認識した結果に基づいて、ある段落Ｘの末尾の文章と文法的に接続可能な文頭を有する段落を接続の可能性のある段落としてもよい。また、レイアウトに特定の規則がある文章については、その規則に基づいた接続の可能性のある段落を選択することでもよい。

【００４８】

図１８のような認識結果（文書データ）から「日本の人口」という文字要素列を検索する場合、第１から第５までの実施の形態と同様に、「日本の人口」を構成する各文字要素「日」、「本」、「の」、「人」、「口」を各段落から順次検索していく。図１８の場合段落Ａの末尾に文字要素「日」、「本」、「の」が続けて検出できる。次に「人」を検索する。ここで、段落Ａに接続する可能性のある段落はＢとＣであるため、段落Ｂと段落Ｃのそれぞれの文頭に「人」が存在するかを判断する。この場合段落Ｂの文頭に「人」が検出できるので、その次の位置に「口」が存在するかを判断する。最終的に「日本の人口」を構成する全ての文字要素が検出できたことになる。

【００４９】

このように接続する可能性のある段落を複数保持しておくことで、文書認識の際に段落の接続を誤判断した場合にでも、複数の段落にまたがる文字要素列を検出することが可能となる。

【００５０】

また、段落同士の接続に限らず、段落内で行と行との接続があいまいな場合（行間に図、表、見出しなどが挿入される場合）にも同様に行ごとに異なる番号を付与し、接続する可能性のある行の番号を行ごとに複数保持しておくことで、複数の行にまたがる文字要素列を正しく検出することが可能となる。また、文字要

素と文字要素との接続があいまいな場合（図、表の挿入が間にある場合や、文字要素列の配置が装飾的な場合[曲線状に配置された文字要素列]など）にも同様に行（文字要素）ごとに異なる番号を付与し、接続する可能性のある行（文字要素）の番号を行（文字要素）ごとに複数保持しておくことで、複数の行（文字要素）にまたがる文字要素列を正しく検出することが可能となる。なお、各段落（行または文字要素）についてその段落（行または文字要素）の前に接続する段落（行または文字要素）の番号を合せて保持する形態でも同様の効果が得られる。また、接続する段落（行または文字要素）の番号の表現としては上記のように段落（行または文字要素）番号の絶対値で表現する以外に、段落（行または文字要素）番号の相対値（段落Ａに接続する段落を段落Ｂ，段落Ｃと表現する代わりに、段落＋１、段落＋２と表現する）で表現しても良い。

【００５１】

（第７の実施の形態）

次に、本発明の第７の実施の形態について説明する。第６の実施の形態と同様に、図１７のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。ここで、文字認識した結果を各段落ごとに付与した固有の番号（図１７ではＡ，Ｂ，Ｃ，Ｄ）および各段落の文書内での位置座標（図１７では右上を原点とし左方向をＸ座標、下方向をＹ座標とする）と共に格納しておく（図１９）。図１９では、段落Ａの位置座標（Ｘ、Ｙ）が（１０、１００）であることを示している。

【００５２】

検索する手続きは第６の実施の形態とほぼ同様であるが、段落Ａの末尾に文字要素「日」、「本」、「の」が続けて検出できた後に「人」を検索する際、段落Ａに接続する可能性のある段落を図１９に格納している段落の座標値を用いて決定する。この場合、段落Ａの座標値は（Ｘ、Ｙ）＝（１０、１００）であるので、隣接する段落の座標としてＸ座標値が等しく、Ｙ座標値が次に大きい段落Ｃ（Ｘ、Ｙ）＝（１０、２００）とＹ座標値が等しくてＸ座標値が次に大きい段落Ｂ（Ｘ、Ｙ）＝（１００、１００）を接続する可能性のある段落と判断し、段落Ｂと段落Ｃのそれぞれの文頭に「人」が存在するかを判断する。この場合段落Ｂの

文頭に「人」が検出できるので、その次の位置に「口」が存在するかを判断する。最終的に「日本の人口」を構成する全ての文字要素が検出できたことになる。

【００５３】

接続する段落を決定する方法としては、上記の例以外にも、縦書きの場合、ある段落Ｘの次に続く段落としては段落Ｘよりも下に位置するか、左に位置する段落を接続の可能性のある段落としても良い。また、レイアウトについて特定の規則がある文章については、その規則に基づいた接続の可能性のある段落を位置座標を用いて選択する。なお、位置座標の表現としては、原点や座標軸は自由に選んで良く、また座標値についても段落や図ごとに番号を割り振った値の順番を座標値の単位として用いてもよい。

【００５４】

このように段落ごとに段落の位置の情報を保持しておくことで、文書認識の際に段落の接続を誤判断した場合にでも、接続する可能性のある段落を選択し、複数の段落にまたがる文字要素列を検出することが可能となる。また、位置座標を保持しておくことで、文書データを変更することなく接続する段落の決定方法を変更することが可能であり、段落の位置座標は文書のレイアウトを再現する為に用いることも可能である。

【００５５】

なお、上記の例では段落ごとに位置座標を保持したが、行単位や文字要素単位に異なる番号を付与し、位置座標と共に格納して、接続する可能性のある行または文字要素を決定して検索することで、複数の行または文字要素にまたがる文字要素列を検索することが可能となる。

【００５６】

（第８の実施の形態）

次に、本発明の第８の実施の形態について説明する。第６の実施の形態と同様に、図１７のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。ここで、文字認識した結果は図２０のように接続する可能性のある段落との可能性を含めた複数の認識結果として文字認識結果（文書データ）に保持しておく。図２０では文字認識結果とし

て２種類保持しており、段落Ａに段落Ｂが接続した場合（文字認識結果２）と段落Ａに段落Ｃが接続した場合（文字認識結果１）とを保持している。図２０の中から「日本の人口」を検索する場合、文字認識結果１と２それぞれに対して「日本の人口」を検索し、文字認識結果２から「日本の人口」を検出することができる。なお、図２０のように複数の段落の接続を想定して認識結果を保持する場合、検索する文字要素数に上限を設け（例えば１０文字要素）、段落Ａに接続する段落Ｂ、段落Ｃの認識結果は段落の文頭から９文字要素のみを段落Ａの認識結果に付加して保持するようにしてもよい。この場合段落Ａから段落Ｂまたは段落Ｃにまたがる１０文字要素までの文字要素列の検索が可能となる。

【００５７】

このように、あらかじめ接続する可能性のある段落を含めて認識結果を複数保持しておくことで、文書認識の際に段落の接続を誤判断する場合にでも、複数の段落にまたがる文字要素列を検出することが可能となる。また、段落間の接続を含めて複数の認識結果を文書データに保持しておくことで、検索手続きは簡易になり、従来法の検索手続きを利用することが可能となる。

【００５８】

（第９の実施の形態）

次に、本発明の第９の実施の形態について説明する。図２１のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「神戸」を検索する場合を考える。図２１のような文書の場合、本来の文書の向きは横書きであるが、文字同士の間隔は縦方向に接近しており、文字認識を行った場合に誤判断する可能性がある。ここで、各文字要素を文字認識した場合に、縦書きを想定した場合と横書きを想定した場合の２種類のレイアウトに対応した認識結果（図２２のａ、ｂ）を作成し、認識結果の文書データとして保持しておく。ここで、「神戸」という文字要素列を縦書き、横書きそれぞれに対応した認識結果から検索を行い、横書きを想定した図２２（ａ）の３行目から「神戸」を検出し、図２１の文書に「神戸」が含まれることが分かる。

【００５９】

このように、複数のレイアウトに対応した認識結果を文書データに保持してお

くことで、レイアウトの判断が困難な文書に対しても認識結果に基づいた文字要素列の検索が可能となる。また、検索手続きは従来法の検索手続きを利用することが可能となる。

【００６０】

なお、上記の例では縦書きと横書きの２種類のレイアウトを想定したが、縦書き・横書き以外にも斜め方向のレイアウトなどその他のレイアウトも同様に扱うことが可能である。

【００６１】

（第１０の実施の形態）

次に、本発明の第１０の実施の形態について説明する。第１０の実施の形態ではレイアウトの判断を誤った認識結果の文書データから指定した文字要素列を検出する手続きを説明する。図２１のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「神戸」を検索する場合を考える。ここで、縦書きを想定した認識結果図２２（ａ）を認識結果（文書データ）として保持していたとする。ここで、「神戸」という文字要素列を構成する文字要素「神」と「戸」とに分けて、それぞれの文字要素について図２２（ａ）から検索を行う。各文字要素の検索を行うと「神」は行番号５の第３文字目に検出でき、「戸」は行番号４の第３文字目に検出できる。ここで、構成する文字要素「神」、「戸」が全て検出できた場合、それぞれの文字要素を検出した位置関係に基づいて、文字要素列「神戸」の検出を判断する。この場合「神」と「戸」が隣接する行の同じ文字数目に連続して検出できたため、文字要素列「神戸」が検出できたとする。個々の文字要素を検出した位置関係としては上記以外の基準を設けても良い。例えば、文字の位置座標が分かっている場合には、個々の文字要素があらかじめ定めた距離以下で接近しかつ直線的に配置してあることを判断基準としても良い。また、検索の手続きとしては他の手続きでも良く、上記のように文字要素を全て検索するのではなく、「神」を検出できた場合にのみ「神」を検出した行に隣接する行からのみ「戸」を検索するようにしても良い。これにより不用な検索手続きを削減し、効率的な文字要素列の検索が可能となる。

【００６２】

このように、検索したい文字要素列を構成する文字要素を個別に文書データから検索し、個々の検出位置の位置関係から文字要素列の有無を判断することで、レイアウトの判断を誤って認識した文書データからでも指定した文字要素列を正しく検索することが可能となる。

〔００６３〕

（第１１の実施の形態）

次に、本発明の第１１の実施の形態について説明する。第１１の実施の形態では段落同士の位置関係を認識結果と共に文書データに保持しておき指定した文字要素列を検出する手続きを説明する。ここで、図１７のようなレイアウトの文章を文字認識した結果（文書データ）から文字要素列「日本の人口」を検索する場合を考える。第６の実施例と同様に、文字認識した結果を各段落ごとに付与した固有の番号（図１７ではＡ，Ｂ，Ｃ，Ｄ）および各段落の文書内での位置座標（図１７では右上を原点とし左方向をＸ座標、下方向をＹ座標とする）と共に格納しておく（図１９）。

〔００６４〕

最初に、指定した文字要素列「日本の人口」を途中で分割し、２つの文字要素列に分ける（例えば「日本」と「の人口」など）。次に分割してできた２つの文字要素列それぞれを個々の段落から検索する。全ての分割の仕方について同様の手続きで検索を行う。「日本の」と「人口」に分割した場合、図１７では段落Ａの末尾に「日本の」が検出でき、段落Ｂの文頭に「人口」が検出できる。分割した文字要素列が全て検出できた場合、検出できた段落同士の位置関係に基づいて、文字要素列「日本の人口」の検出を判断する。例えば、２つの文字要素列を検出した段落が隣接していたり、位置が近い場合には指定した文字要素列を検出できたとする。図１７の場合、「日本の」が検出できた段落Ａの位置座標（ｘ、ｙ）＝（１０，１００）と「人口」が検出できた段落Ｂの位置座標（ｘ、ｙ）＝（１００，１００）は同じｙ座標で隣接することから「日本の人口」が検出できたとする。また、上記のように文字要素列を分割して得た文字要素列を段落内から検索する場合には、各段落の文末と文頭のみに検索処理を行えば検索効率が良い。なお、上記の例は指定した文字要素列を２つに分割したが必要に応じて３つ以

上に分割しても同様の手続きが可能である。

【００６５】

また、上記の例では複数の段落にまたがる文字要素列を検索する例を示したが、複数の行にまたがる文字要素列についても同様に検索が可能である。この場合指定した文字要素列を分割し、各行に対して分割した文字要素列それぞれを検索し、分割した全ての文字要素列が隣接して検出できた場合にもとの指定した文字要素列が検出できたとする。

【００６６】

このように本実施例では、段落や行の接続が誤っている（または不定な）場合にでも、複数の段落にまたがる文字要素列を正しく検出することが可能となる。

【００６７】

このように本発明では、文字認識の誤りがある場合や、段落間・行間の接続が誤っていたり不定な場合や、縦書き・横書きの判断が誤っているあるいは不定である場合に指定した文字要素列を検索することが可能である。

【００６８】

なお、本発明の第１から第１１の実施の形態は単独で用いてもよいし、組み合わせて実施することも可能である。また、上記実施の形態の実現手段としてはハードウェアを用いて実現してもよいし、あるいはコンピュータ上のソフトウェアを用いて実現してもよい。

【００６９】

また、上記実施の形態の各手続きのうち、全部または一部の手続きをコンピュータに実行させるためのプログラムを記録した媒体を用いる、あるいは通信網（ネットワーク）または放送を通じてプログラム（又はその一部）をダウンロードして実行することでも、上記の場合と同様の効果を実現することが可能である。

【００７０】

（第１２の実施の形態）

次に本発明の第１２の実施の形態について説明する。本実施の形態は、検索語を構成する文字要素の要素数、および、検索対象である文書データ群中の文字要素から検出された検索語の文字要素数別に、検索した文字要素列が検索対象の文

字要素列であるかを表す確率を記述したテーブルを用いた検索の例である。以下は、文字要素を、単に文字として表現する。通常、さまざまな言語の単語には冗長性があり、数文字が分からない場合でも単語が特定できる場合が多い。この傾向は、単語を構成する文字数が多ければ多いほど当てはまる。本実施の形態は、単語のこのような傾向を利用して誤りを含んだ文字列から単語を検索できることを示す。

【００７１】

ここでは、図２６のように、本来「・・・オックスフォード大学は・・・」という文字画像列を含む文書画像を文字認識して得られた「・・・オッタスフォード大学は・・・」という認識結果（文書データ、予め記憶媒体などに保持しておく）から検索語「オックスフォード」を検索することを考える。各認識結果には、認識結果の確からしさ（正解確率）を表す信頼度が付与されている。

【００７２】

検索を行う前に、予め、図２７のような確率テーブルを算出しておく。図２７は、検索語を構成する文字数がｎ文字のとき、検索対象の文書中の文字列とｋ文字が一致した場合に、その部分が検索語と一致する確率Pa(n,k)を表したテーブルである。この確率テーブルは、誤りを含まない大量のテキストデータと単語辞書により算出したものである。算出方法としては、まず、単語辞書中の単語（ｎ文字）について、テキストデータ中の全ての連続するｎ文字に対して、単語と一致する文字数を調べ、一致する文字数別に累計Nk(i=1, ...,n)をとる。この累計Nkを用いると、ｎ文字の検索語のうちｋ文字が一致した場合に、その部分が検索語である確率Pa(n,k)（＝Nk／Nn）が算出できる。

【００７３】

この確率Pa(n,k)は、文字数ｎが一致していても単語の表記が異なる場合や、一致する文字数ｋの位置によって、当然異なるものとなるが、本実施の形態では、単語の表記や、一致する文字の位置に関しては区別をしていない。すなわち、文字数ｎが等しい全ての単語について、ｋ文字が一致する回数をそれぞれ累計し、その累計の和（平均でもよい）を用いて算出している。

【００７４】

なお、単語の表記別や、一致した文字の位置別、単語を構成する字種別（漢字、ひらがな、カタカナ、英字など）にこのような確率を算出しておいて使用してもよい。

【００７５】

なお、単語の文字数が多い場合は、一致した文字数ｋがある程度になると、単語の文字数ｎに関わらず一致した文字数ｋだけで確率が表せる（Pa(n,k)において、ｎが変化してもほぼ一定）場合がある。この場合は、一致した文字数ｋのみに依存する確率Pa(k)をPa(n,k)の代わりに用いてもよい。

【００７６】

次は、実際に検索を行う場合であるが、図２６の場合だと、検索語「オックスフォード」に対し、検索対象の文書においては、誤認識の「タ」（信頼度0.42）を除く全ての文字が一致する。

【００７７】

ここで、この照合箇所の正当性を表すPwを（数１）で表す。

【００７８】

【数１】

$$P_w = P_a(n, k) \cdot P_b(k)$$

【００７９】

（数１）において、Pa(n,k)は、ｎ文字の単語のうちｋ文字が一致した場合に、その部分が検索語である確率であり、この場合は「オックスフォード」というｎ＝８文字の単語のうちｋ＝７文字が一致しているため、図２７よりPa(8,7)=0.9となる。

【００８０】

また、Pb(k)は、ｋ文字が全て検索対象の文字である確率を表す。各文字に付与されている信頼度は確率であるから、認識結果と一致した文字の信頼度の積とすればよい。図２７よりPb(7)=0.98×0.97×0.99×0.98×0.99×0.97×0.96=0.85となる。よって、この照合箇所の正当性を表す値はPw=Pa(8,7)×Pb(7)=0.9×0.

85=0.765となり、予め定めた閾値(本実施の形態では0.6とする)よりも大きいため、この照合箇所を検索結果とする。

【００８１】

なお、長い文字数を持った単語の場合は、信頼度の積Pb(k)が小さくなりやすいため、何らかの方法で文字数に対して正規化してもよい。また、各文字の信頼度が確率でない場合は、確率に変換して用いたり、単純に平均を求めてPb(k)の代わりにしてもよい。

【００８２】

また、図２８のように、正解した文字が７文字であっても低い信頼度を持った文字が存在する場合は、その文字をカウントしない場合が結果的にPwとしては大きくなる場合があるため、Pwが最大となるような正解文字数kを選んでもよい（７文字正解とすると Pw=Pa(8,7)×Pb(7)= 0.90×(0.98×0.97×0.99×0.98×0.99×0.97×0.30)=0.239 だが、８文字目の正解を含めないで6文字正解とするとPw=Pa(8,6)xPb(6)=0.85×(0.98×0.97×0.99×0.98×0.99×0.97)=0.752であるため後者をPwとして採用する）。

【００８３】

なお、文字毎に信頼度が付与されていないデータベースの場合は、検索語の長さｎと、一致した文字数ｋを用いてPa(n,k)を、照合した箇所の正当性を表す値としてもよい。

【００８４】

なお、一致しなかった文字は今回情報として用いていないが、一致しなかった文字の信頼度が高い場合はその文字が正解である可能性が高い（すなわち、１文字だけ異なる単語が存在し、検索結果としては正当性が低い）と考えられるので、一致しない文字の信頼度を利用したペナルティを導入してもよい（一致しない文字の信頼度が予め定めた閾値よりも高い、または、そのような文字が予め定めた文字数よりも多い場合は、照合箇所の正当性を表すPwが閾値より大きくても検索結果として採用しないなど）。

【００８５】

このように、単語の冗長性を利用して、全ての文字が一致しなくても誤認識を

含んだテキスト文書中から検索が可能となる。また、一致しない文字の個数をどう決めたらよいかという問題も、実際の大量のテキストデータベースから図２７のような確率テーブルを用いて（数１）の式で検索箇所の正当性を数値化することによって解決することができる。

【００８６】

（第１３の実施の形態）

次に本発明の第１３の実施の形態について説明する。本実施の形態は、第１２の実施の形態で説明した検索語を構成する文字要素の要素数、および、検索対象である文書データ群中の文字要素から検出された検索語の文字要素数別に、検索した文字要素列が検索対象の文字要素列であるかを表す確率を記述したテーブル、および、第３の実施の形態などで説明した文字要素同士の距離を示す距離テーブルを用いた検索の例である。以下は、文字要素を、単に文字として表現する。

【００８７】

ここでは、図２９のように、本来「・・・オックスフォード大学は・・・」という文字画像列を含む文書画像を文字認識して得られた「・・・オッタヌフォード大学は・・・」という認識結果（文書データ、予め記憶媒体などに保持しておく）から検索語「オックスフォード」を検索することを考える。各認識結果には、認識結果の確からしさ（正解確率）を表す信頼度が付与されている。第１２の実施の形態と同様、検索を行う前に図２７のような確率のテーブルを算出しておく。

【００８８】

検索時には、第１の実施の形態と同様に、各文字に付与された信頼度を文字要素間距離テーブルを参照する基準距離に変換し、第３の実施の形態と同様に、距離を発生頻度（確率）に変換する（図２９）。

【００８９】

そして、この照合箇所の正当性を表すPw（数２）を図２７のテーブルから算出する。

【００９０】

【数２】

$$P_w = P_a(n, k)$$

【００９１】

認識結果そのものが一致した文字数は１，２，５，６，７，８文字目の６文字であるが、３文字目に関しては文字要素間テーブルを参照することにより、結果的に７文字一致していることになる（信頼度0.42の認識結果「タ」は、「ク」である確率（頻度）が0.3 > 0である）。よって、Pw=Pa(8,7)=0.9となる。ここで、予め定めた閾値を0.80とすると、この閾値より大きいということで、検索結果としてもよいが、検索の目的によっては検索したい文字列以外の文字列が検索されてしまう“検索ノイズ”を出来るだけ減らしたいということもある。その場合は、より詳細な正当性の判定として、一致しなかった４文字目の「ヌ」について、距離の基準値を大きくする（すなわち認識信頼度を小さく設定しなおし、距離の基準値を求める）ことにより、距離が２０のときには、検出できなかった（頻度０）認識誤りを検出することも可能である。

【００９２】

一致しなかった文字に対して、全ての文字の可能性を許容したワイルドカードの扱いにすると、偶然１文字だけ異なる別の単語が検索されてしまう可能性が大きいが、このように、少し距離の基準値を大きくすることによって、文字要素間テーブルにおいて誤認識しやすい類似文字のみ、誤認識を想定して文字の検索を行うことにより、検索ノイズを減らすことができる。

【００９３】

また、距離の基準値を大きくするということは、誤認識で認識信頼度が大きい場合に発生頻度が０（信頼度が大きいがために距離の基準値が小さめになるが、もう少し基準値が大きければ頻度＞０）となる場合を救済することができる。

【００９４】

なお、正当性を表すPwの値によって、距離の基準値を大きくする幅（認識信頼度の下げ幅）を制御してもよい。また、そのような距離の基準値を大きくする（

認識信頼度を小さくする）文字数を制御してもよい。

【００９５】

なお、正当性を表すPwの値によって、距離の基準値を大きくするか、ワイルドカード扱いにするかを制御してもよい。

【００９６】

（第１４の実施の形態）

次に本発明の第１４の実施の形態について説明する。本実施の形態は、第１２、第１３の実施の形態で説明した検索語を構成する文字要素の要素数、および、検索対象である文書データ群中の文字要素から検出された検索語の文字要素数別に正解確率を記述したテーブルを用いた検索の例である。以下は、文字要素を、単に文字として表現する。

【００９７】

ここでは、図３０のように、本来「・・・オックスフォードの学生達・・・」という文字画像列を含む文書画像を文字認識して得られた「・・・オッタスフォード○学生達・・・」という認識結果（文書データ、予め記憶媒体などに保持しておく）から検索語「オックスフォード大学」を検索することを考える。各認識結果には、認識結果の確からしさ（正解確率）を表す信頼度が付与されている。第１２、第１３の実施の形態と同様、検索を行う前に図２７のような確率のテーブルを算出しておく。

【００９８】

検索時には、まず検索語「オックスフォード大学」を、予め用意した単語辞書を用いて単語の組み合わせに分割できるかどうか調べる。本実施の形態で用意した単語辞書には、「オックスフォード」「大学」という単語が存在しており、検索語「オックスフォード大学」を「オックスフォード」＋「大学」と分割する。そして検索時には「オックスフォード」という検索語の後に、「大学」という検索語がある場所を検索対象の文書データの中から探す。それぞれの探し方は、第１２または第１３の実施の形態と同様である。

【００９９】

もし、単語に分割しないで検索した場合、図３０の例では「オックスフォード

大学」という１０文字の単語において１，２，４，５，６，７，８，１０文字目の８文字が一致する。また、通常Pa(10,8)は非常に大きいため、誤検出されて検索ノイズとなり易い。他の場所でも、単語「オックスフォード」が存在する近辺では検索がヒットし易くなり、検索ノイズが頻発してしまうという恐れがある。このように、複数の単語で長い検索語が形成されている場合は、検索語を、単語辞書などを用いて複数の単語に分割することで、検索ノイズを低減することが出来る。

【０１００】

なお、単語辞書には、通常の単語だけでなく、複数の単語に共通して表れ易い部分文字列も含んでもよい。例えば、「プランテーション」「オリエンテーション」「ステーション」などに含まれる「テーション」という文字列を辞書に含んでおき、検索時には、それぞれの文字列を「プラン」＋「テーション」、あるいは、「オリエン」＋「テーション」、あるいは、「ス」＋「テーション」と分割してそれぞれ検索することによって、検索語の一部が等しい別の単語が誤検出されるのを防ぐことが出来る。

【０１０１】

【発明の効果】

以上のように本発明は、文字要素（文字または文字片または文字列または文字片と文字との組合わせ）同士の距離を表現したテーブルを用いることで、認識結果に対して許容できる誤り度合いを動的に変更して検索を行うことが可能である。また、テーブルを用いることで、複雑な距離計算を行わず高速な検索が可能である。

【０１０２】

また、文字要素同士の接続関係を複数保持する、あるいは複数通り検索する、あるいは文字要素列を分割して検索することで、文書のレイアウトを誤って解釈した文書データから望ましい検索を実現することが可能である。これにより、縦書き・横書きを間違って判断している文書データや、改行後に継続する行を誤って判断している文書データからの文字列検索が可能となる。

【図面の簡単な説明】

【図１】

本発明の第１から第４の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図２】

文字片の例を示す図

【図３】

文字片と文字とが集まった文字要素の例を示す図

【図４】

複数の文字要素を単一の文字として扱った結果生じる誤認識の例を示す図

【図５】

単一の文字を複数の文字として扱った結果生じる誤認識の例を示す図

【図６】

第２の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図７】

検索する文書の例と認識結果の一例を示す図

【図８】

第２の実施の形態で用いる明朝体文字の文字要素間距離テーブルの一例を示す図

【図９】

第１の実施の形態で検索パラメータと基準距離の対応の一例を示す図

【図１０】

第１の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図１１】

第３の実施の形態で検索パラメータと基準距離、発生頻度との対応の一例を示す図

【図１２】

第３の実施の形態で検索パラメータと基準距離、発生頻度との対応の一例を示す図

【図１３】

第３の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図１４】

第３の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図１５】

第３の実施の形態で用いる文字要素間距離テーブルの一例を示す図

【図１６】

第５の実施の形態で用いる複数の文字候補を保持した文書データの一例を示す図

【図１７】

検索する文書の例を示す図

【図１８】

第６の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図１９】

第７、第１１の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図２０】

第８の実施の形態で用いる認識結果を保持した文書データの一例を示す図

【図２１】

検索する文書の例を示す図

【図２２】

第９、１０の実施の形態で用いる認識結果を保持した文書データの一例を示す

（ａ）縦書きを示す図

（ｂ）横書きを示す図

【図２３】

検索する文書の例と認識結果の例を示す図

【図２４】

検索する文書の例を示す図

【図２５】

検索する文書の例と認識結果の例を示す図

【図２６】

第１２の実施の形態で用いる認識結果を保持した文書データの例を示す図

【図２７】

第１２，１３の実施の形態で用いる、単語の文字数と検索時に一致した文字数による、その部分が検索語と一致する確率テーブルの例を示す図

【図２８】

第１２の実施の形態で用いる認識結果を保持した文書データの例を示す図

【図２９】

第１３の実施の形態で用いる認識結果を保持した文書データの例を示す図

【図３０】

第１４の実施の形態で用いる認識結果を保持した文書データの例を示す図

【書類名】　　図面

【図１】

	亜	唖	彡	卩	００
亜		１０	１３２	１６６	１７２
唖			１１５	１５２	１６４
彡				１４３	１９１
卩					６９
００					

【図２】

「彡」、「卩」

【図３】

「）彡」、「卩１」

【図４】

「木」、「木」→「林」
「０」、「０」→「∞」

【図５】

「川」→「１」、「１」、「１」
「い」→「し」、「１」

【図６】

	林	∞	し１	１１１	川
木木	１０	２２１	１９０	１５６	１５２
川	１５５	１６５	９１	９	
い	２０１	１１９	１３	８９	９５
卩	１４９	１８８	９８	１３３	１３７
００	２１５	１２	１０５	１６９	１７２

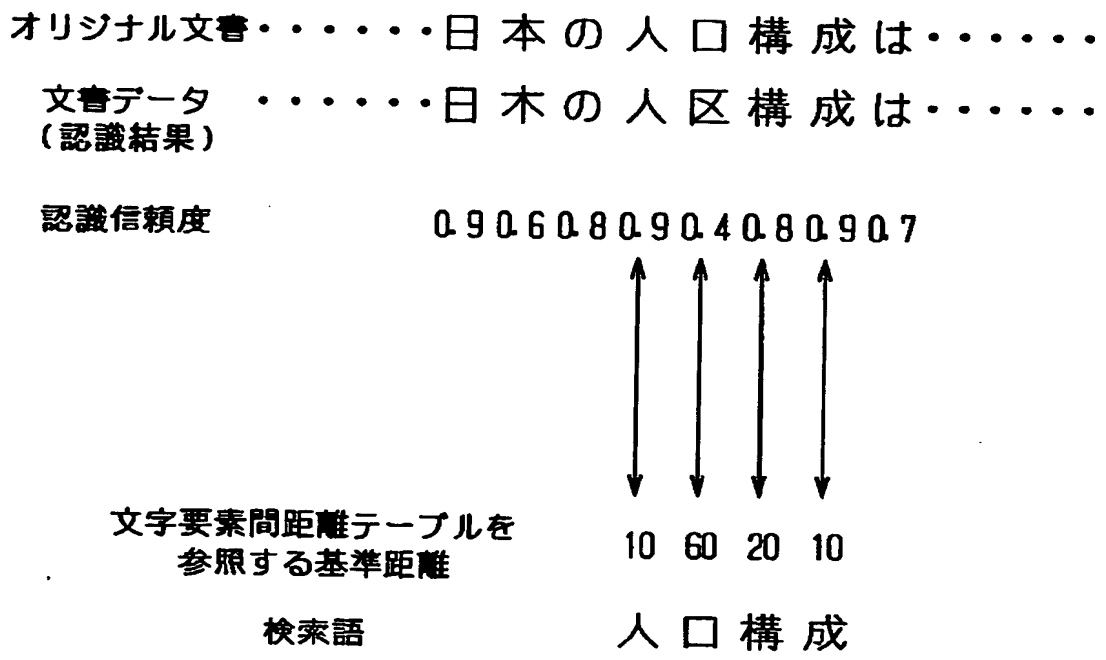
【図７】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日本の人区構成は・・・

【図８】

	亜	唖	彡	卩	００
亜		１２	１３０	１７０	１６８
唖			１１４	１５０	１７０
彡				１４７	１９０
卩					６０
００					

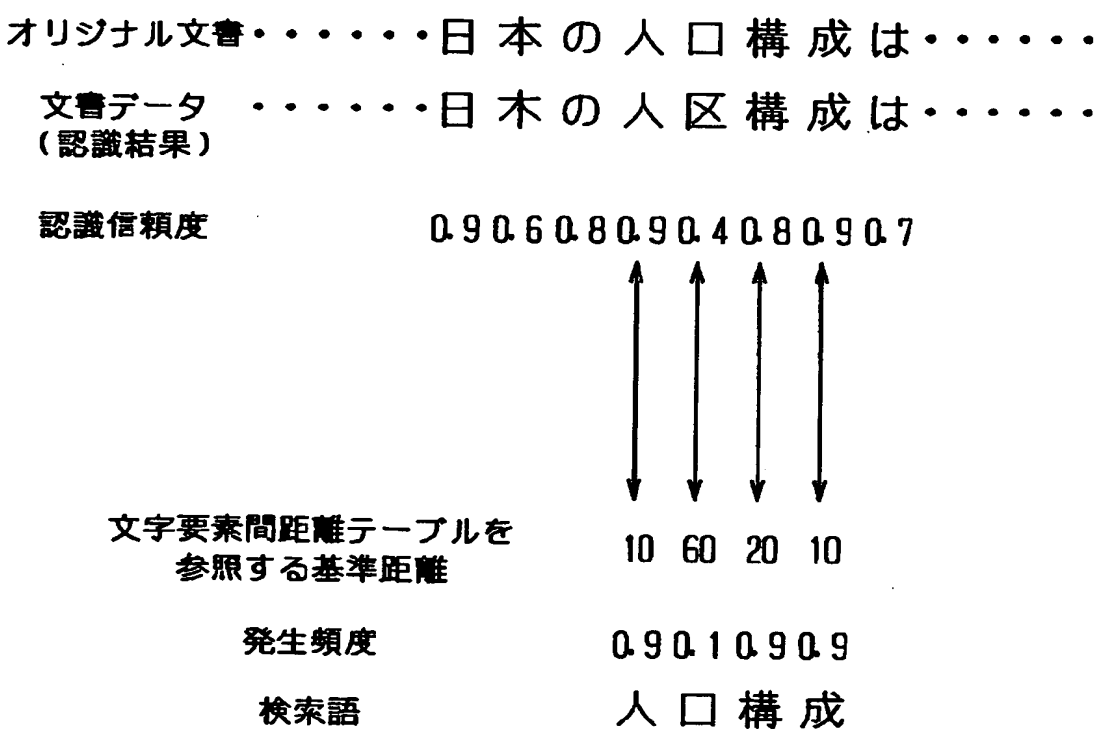
【図９】



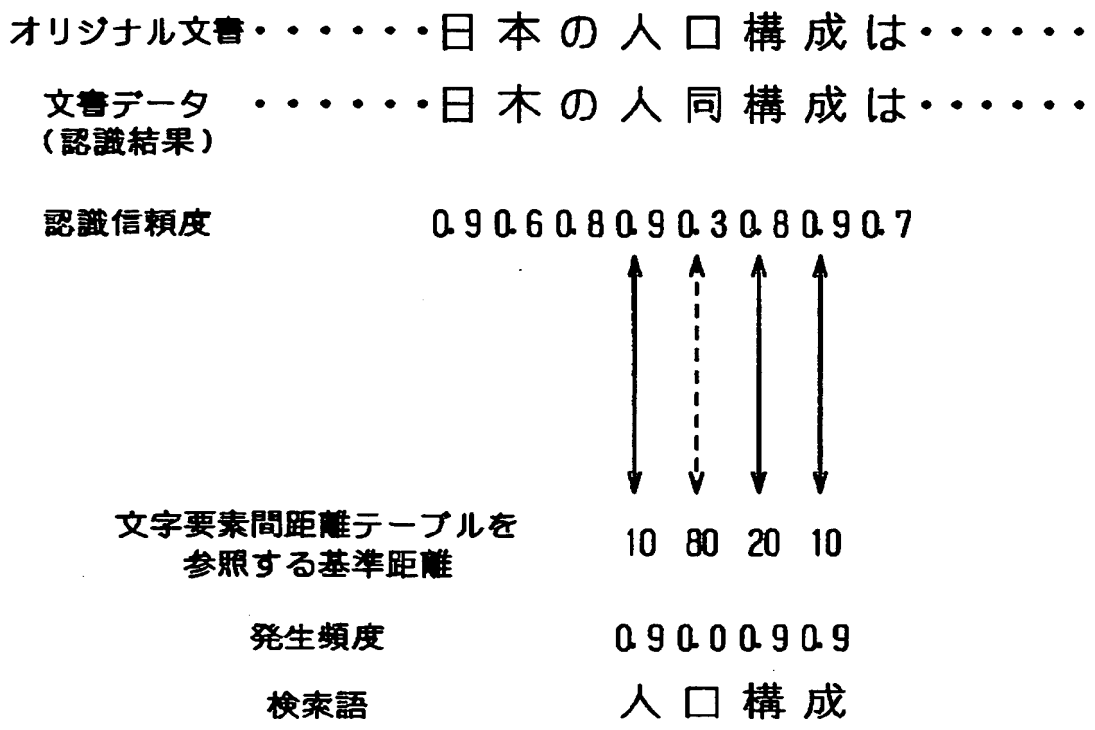
【図１０】

	人	口	構	成	区	同
人	0	170	250	210	99	113
口	170	0	244	168	50	100
構	250	244	0	142	198	184
成	210	168	142	0	137	152

【図１１】


オリジナル文書・・・・・・日本の人口構成は・・・・・・
文書データ（認識結果）・・・・・・日本の人区構成は・・・・・・
認識信頼度
0.9 0.6 0.8 0.9 0.4 0.8 0.9 0.7
文字要素間距離テーブルを参照する基準距離
10 60 20 10
発生頻度
0.9 0.1 0.9 0.9
検索語
人口構成

【図１２】



【図１３】

		T		
下	距離	10	20	30
	確率	0.2	0.6	0.2

【図１４】

		T	F	ト	Γ	木
下	距離	20	60	90	125	130
	分散	10	10	30	25	20

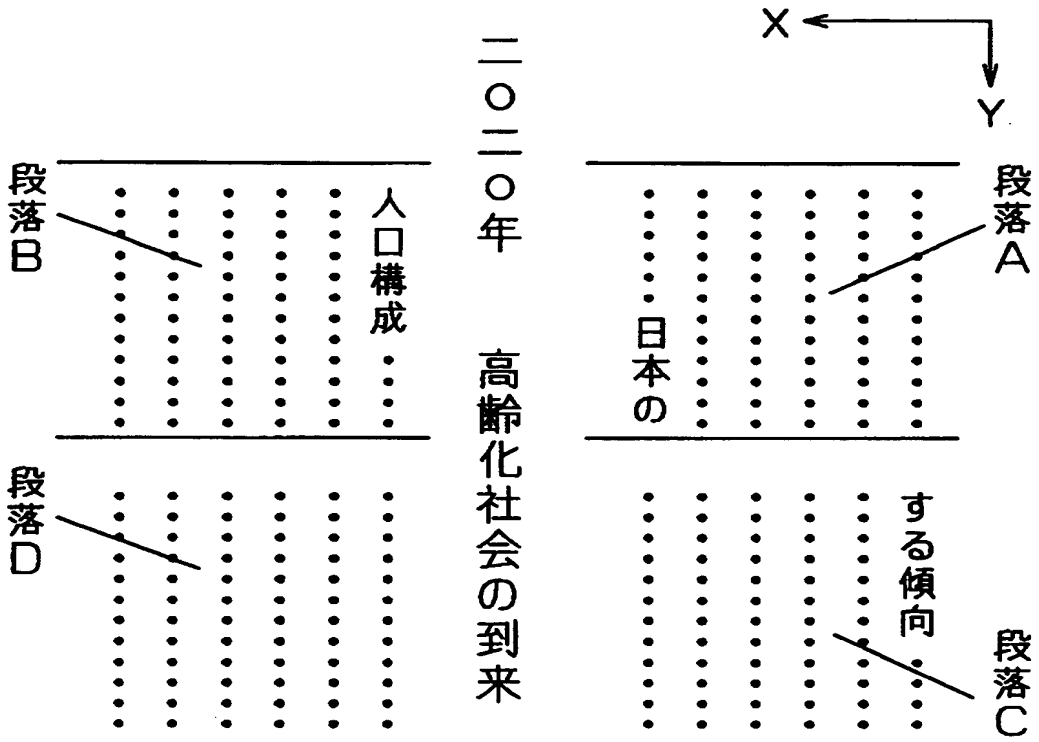
【図１５】

		T	F	ト	Γ	木
下	最短距離	10	50	63	102	110
	最大距離	30	70	122	151	165

【図１６】

認識結果	・・・	日	木	の	人	区	構	成	は	・・・
候補		目	本	@	入	凶	講	茂	ほ	
候補			大		ル	図		感	ぼ	
候補						口				

【図１７】



【図１８】

段落番号	接続する段落の番号	段落単位の認識結果
A	B,C	・・・・・日本の
B	C,D	人口構成は ・・・
C	D	する傾向 ・・・・
D		・・・・・・・・・

【図１９】

段落番号	段落単位の認識結果	段落の位置	
		x	y
A	・・・・・日本の	１０	１００
B	人口構成は ・・・	１００	１００
C	する傾向 ・・・・	１０	２００
D	・・・・・・・・・	１００	２００

【図２０】

文字認識結果１	・・・日本のする傾向 ・・・
２	・・・日本の人口構成は・・・

【図２１】

京　都　　２９℃
大　阪　　３２℃
神　戸　　３０℃

【図２２】

(a)

列番号	列単位の認識結果
1	℃℃℃
2	920
3	233
4	都阪戸
5	京大神

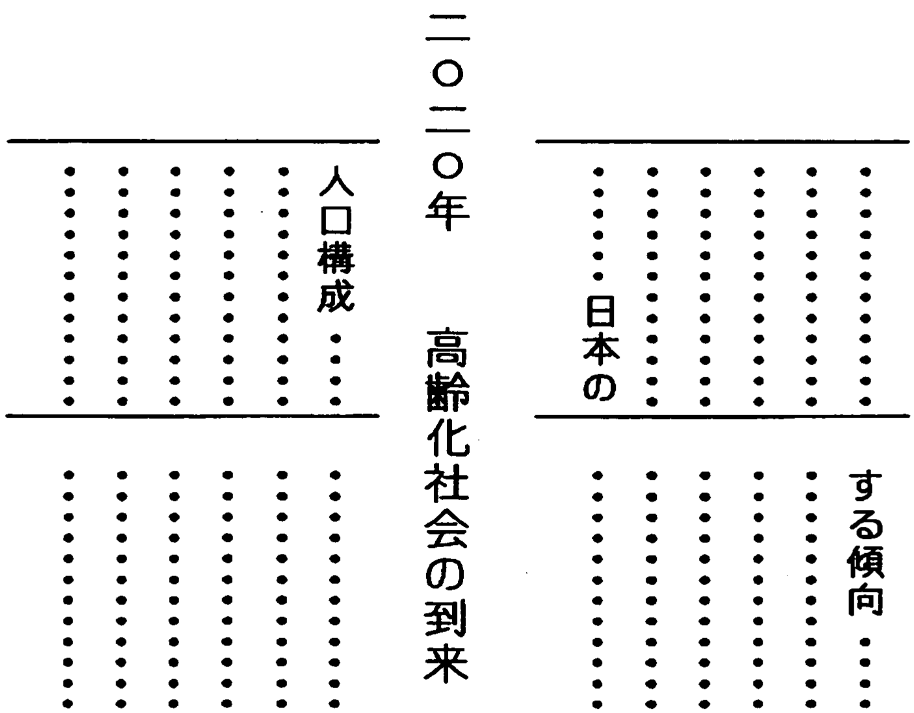
(b)

行番号	行単位の認識結果
1	京都29℃
2	大阪32℃
3	神戸30℃

【図２３】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日木の人区構成は・・・

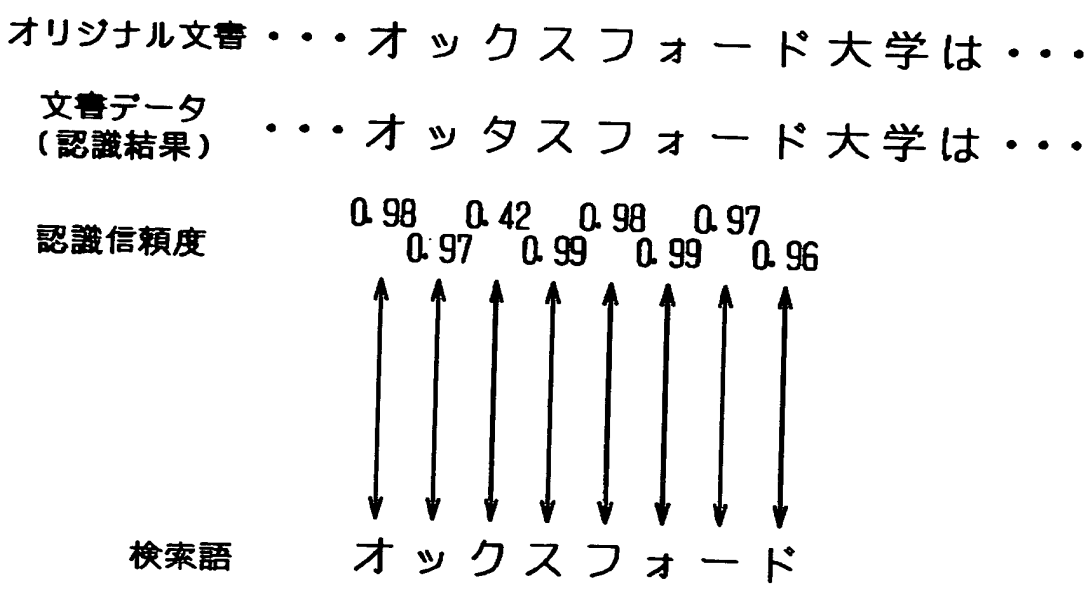
【図２４】



【図２５】

オリジナル文書	・・・日本の人口構成は・・・
文字認識結果	・・・日木のする傾向　・・・

【図２６】



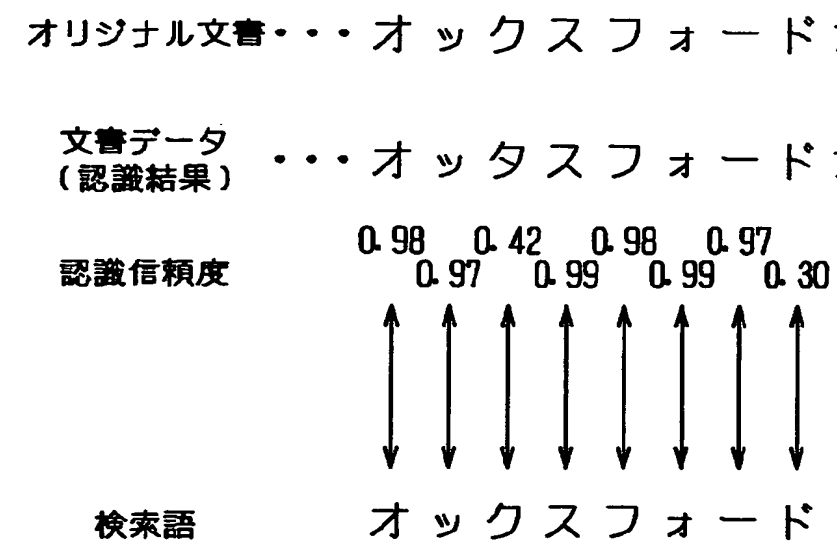
【図２７】

$P_a(n,k)$

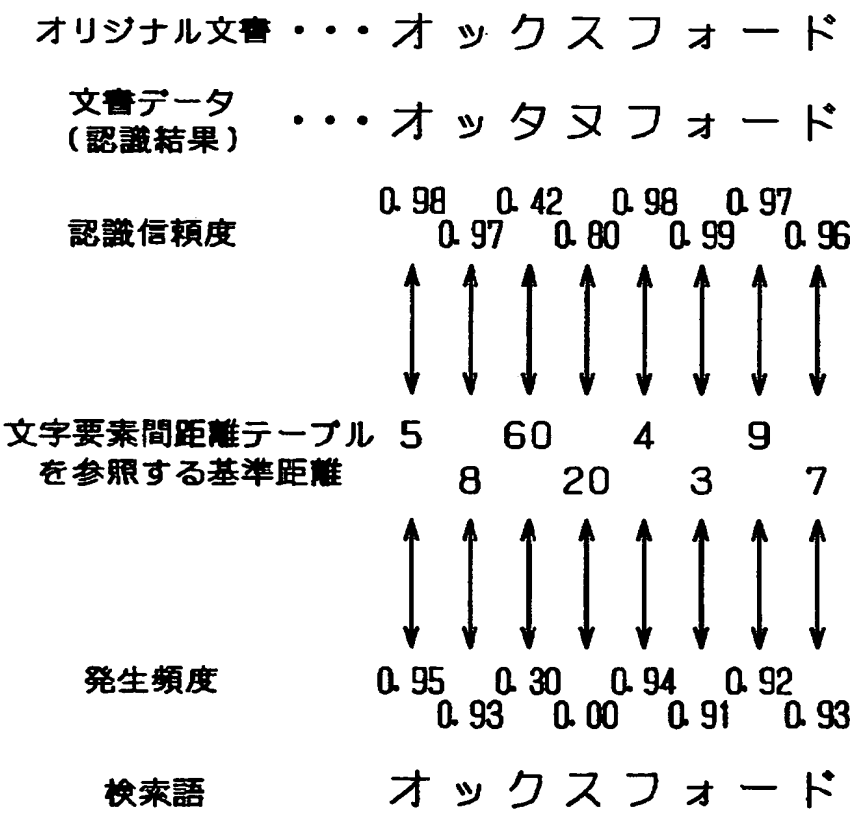
一致文字数k ＼ 単語の文字数n	1	2	3	4	5	6	7	8	・・・
1	1.0	0.1	0.1	0.1	0.08	0.05	0.03	0.01	
2		1.0	0.4	0.2	0.15	0.1	0.05	0.02	
3			1.0	0.6	0.4	0.3	0.2	0.1	
4				1.0	0.8	0.7	0.4	0.4	
5					1.0	0.9	0.8	0.8	
6						1.0	0.9	0.85	
7							1.0	0.9	
8								1.0	

⋮

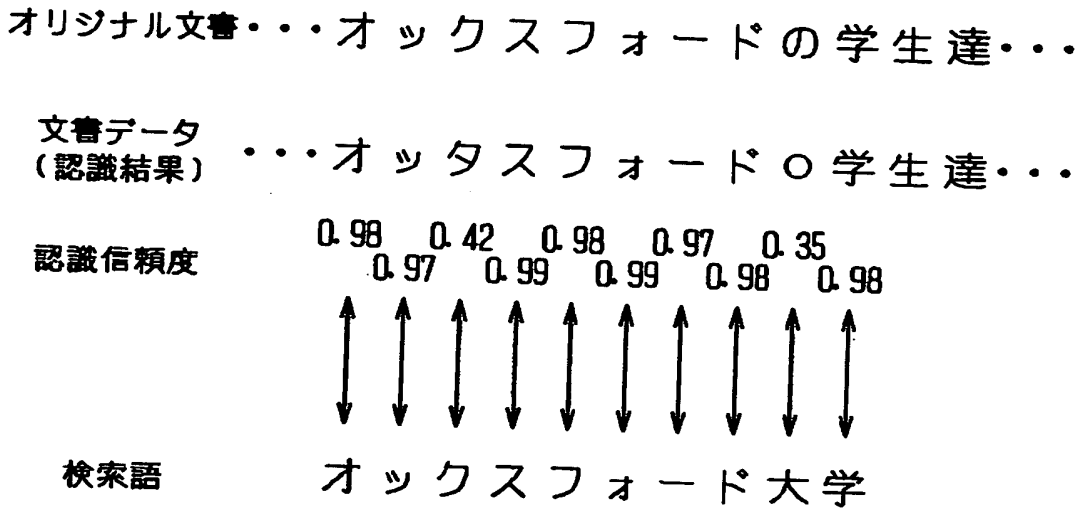
【図２８】


オリジナル文書・・・オックスフォード大学は・・・
文書データ
（認識結果）
・・・オッタスフォード大学は・・・
認識信頼度
0.98 0.42 0.98 0.97
0.97 0.99 0.99 0.30
検索語
オックスフォード

【図２９】


オリジナル文書・・・オックスフォード大学は・・・
文書データ
（認識結果）
・・・オッタヌフォード大学は・・・
認識信頼度
0.98 0.42 0.98 0.97
0.97 0.80 0.99 0.96
文字要素間距離テーブル
を参照する基準距離
5 60 4 9
8 20 3 7
発生頻度
0.95 0.30 0.94 0.92
0.93 0.00 0.91 0.93
検索語
オックスフォード

【図３０】


オリジナル文書・・・オックスフォードの学生達・・・
文書データ
（認識結果）
・・・オッタスフォードＯ学生達・・・
認識信頼度
0.98 0.42 0.98 0.97 0.35
0.97 0.99 0.99 0.98 0.98
検索語
オックスフォード大学

【書類名】　要約書

【要約】

【課題】　文章を文字認識した場合に生じる誤りを含んだ認識結果に対して、指定した文字列を誤認識のない状態で検索する場合と同様に検出すること。

【解決手段】　１つ以上の文字およびまたは１つ以上の文字片から成る単位を文字要素とし、文書データ群の中から、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索する検索処理方法であり、また、文字要素間の距離を表現したテーブルを用いて、指定した文字列を構成する文字要素とあらかじめ定めた関係を満たす文字要素列を検索する検索処理方法である。

【選択図】　図１

出 願 人 履 歴 情 報

識別番号　〔０００００５８２１〕

１．変更年月日　１９９０年　８月２８日
［変更理由］　新規登録
住　所　大阪府門真市大字門真１００６番地
氏　名　松下電器産業株式会社